## WHAT IS CLAIMED IS:

1	1. A method for matching a reference document with a plurality of corpus
2	documents, the method comprising:
3	deriving semantic content of the reference document according to a
4	hierarchical arrangement of semantic types; and
5	for each corpus document,
6	deriving semantic content of the corpus document according to the
7	hierarchical arrangement of semantic types; and
8	producing a matching score for the corpus document by determining a
9	relatedness between the corpus document and the reference document from the derived
O	semantic content of the corpus document and the derived semantic content of the reference
1	document.
1	2. The method recited in claim 1 wherein deriving semantic content of
2	the reference document and deriving semantic content of the corpus document comprises:
3	creating tokenized elements from a text stream;
4	tagging each tokenized element with a grammatical category label; and
5	creating a root form for each tokenized and tagged element.
1	3. The method recited in claim 2 wherein deriving semantic content of
2	the reference document and deriving semantic content of the corpus document further
3	comprises assigning a semantic type within the hierarchical arrangement of semantic types to
4	the root form.
1	4. The method recited in claim 1 wherein producing the matching score
2	comprises determining a distance within the hierarchical arrangement between a semantic
3	type that defines semantic content of the reference document and a semantic type that defines
4	semantic content of the corpus document.
1	5. The method recited in claim 4 wherein determining the distance
2	comprises accounting for a qualia relationship between types in the hierarchical arrangement.
1	6. The method recited in claim 5 wherein the qualia relationship
2	comprises a direct qualia relationship.

4

1

2

- The method recited in claim 5 wherein the qualia relationship 7. 1 comprises an indirect qualia relationship. 2 The method recited in claim 5 wherein the qualia relationship 8. 1 2 comprises a telic relationship. The method recited in claim 5 wherein the qualia relationship 9. 1 2 comprises an agentive relationship. The method recited in claim 4 wherein producing the matching score 1 10. further comprises accounting for whether the semantic type that defines semantic content of 2 the reference document and the semantic type that defines semantic content of the corpus 3 document are in a subsumption relationship. 4 The method recited in claim 4 wherein producing the matching score 11. further comprises applying a filtering function to increase importance of a smaller distance relative to a larger distance. The method recited in claim 11 wherein the filtering function 1 12. comprises a Gaussian function. The method recited in claim 11 wherein the filtering function 13. 1 2 comprises an exponential function. The method recited in claim 11 wherein the filtering function 1 14. 2 comprises a rectangular function. The method recited in claim 1 further comprising ranking the plurality 1 15. of corpus documents in accordance with the matching score for each corpus document. 2 The method recited in claim 1 wherein the plurality of corpus 16. 1 documents is categorized according to a categorization scheme and the reference document 2
  - 17. The method recited in claim 16 wherein the categorization scheme comprises a hierarchical categorization scheme.

comprises an uncategorized document, the method further comprising categorizing the

uncategorized document according to the categorization scheme with the matching score.

1	18. The method recited in claim 17 wherein the plurality of corpus
2	documents is comprised by a larger set of documents within the hierarchical categorization
3	scheme.
1	19. The method recited in claim 1 wherein the reference document
2	comprises a user query.
1	20. The method recited in claim 19 wherein the plurality of corpus
2	documents comprises a plurality of sponsor web pages, the method further comprising
3	generating an output interest statement with semantic structures derived from at least one of
4	the reference document and the corpus document having the highest matching score.
1	21. The method recited in claim 1 wherein the reference document and the
2	plurality of corpus documents are comprised by a document set, the method further
3	comprising:
4	determining the matching scores for a plurality of divisions of the document
5	set into the reference document and the corpus documents;
6	combining the matching scores for each document pair comprised by the
7	document set; and
8	clustering documents within the document set by setting a threshold for the
9	combined matching scores.
1	22. A method for categorizing an uncategorized document within a
2	categorization scheme, the method comprising:
3	deriving semantic content of the reference document according to a
4	hierarchical arrangement of semantic types;
5	performing a comparison of the semantic content of the uncategorized
6	document with semantic content of documents previously categorized according to the
7	categorization scheme; and
8	determining a category for the uncategorized document from the comparison.
1	23. The method recited in claim 22 wherein the categorization scheme
2	comprises a hierarchical categorization scheme.

1	24. The method recited in claim 23 wherein performing the comparison
2	comprises, for each level of the hierarchical categorization scheme:
3	producing a matching score for each unexcluded document categorized at such
4	level; and
5	excluding documents at a level subordinate to such level from the matching
6	score.
1	25. The method recited in claim 22 wherein determining a category for the
2	uncategorized document comprises determining a plurality of categories for the document.
1	26. The method recited in claim 22 wherein performing a comparison
2	comprises producing a matching score for each of the plurality of documents previously
3	categorized by determining a relatedness with the uncategorized document.
1	
1	27. The method recited in claim 26 wherein producing the matching score
2	comprises determining a distance within the hierarchical arrangement between a semantic
3	type that defines content of the uncategorized document and a semantic type that defines
4	semantic content of the previously categorized document.
1	28. The method recited in claim 27 wherein determining the distance
2	comprises accounting for a qualia relationship between types in the hierarchical arrangement.
1	29. The method recited in claim 27 wherein producing the matching score
2	further comprises accounting for whether the semantic type that defines semantic content of
3	the uncategorized document and the semantic type that defines semantic content of the
4	previously categorized document are in a subsumption relationship.
	1
1	30. The method recited in claim 27 wherein producing the matching score
2	further comprises applying a filtering function to increase importance of a smaller distance
3	relative to a larger distance.
1	31. A system for matching a reference document with a plurality of corpus
2	documents, the system comprising:
3	a database configured for storing a hierarchical arrangement of semantic types;
4	and
5	an engine in communication with the database configured to

6	derive semantic content of the reference document and of each corpus
7	document according to the hierarchical arrangement; and
8	produce a matching score between the reference document and each
9	corpus document from the derived semantic content.
1	32. The system recited in claim 31 wherein the engine is further
2	configured to rank each corpus document according to its matching score.
Ī	33. The system recited in claim 31 wherein the engine is configured to
2	produce the matching score by determining a distance within the hierarchical arrangement.
1	34. The system recited in claim 33 wherein determining the distance
2	comprises accounting for a qualia relationship between types in the hierarchical arrangement.
1	35. The system recited in claim 33 wherein the matching score is filtered
2	to increase the importance of a smaller distance relative to a larger distance.
1 2	36. The system recited in claim 31 wherein the engine is in communication with the internet.
2	with the internet.
1	37. A system for categorizing an uncategorized document within a
2	categorization scheme, the system comprising:
3	a database configured for storing a categorization for each of a plurality of
4	previously categorized documents and for storing a hierarchical arrangement of semantic
5	types; and
6	an engine in communication with the database configured to
7	derive semantic content of the uncategorized document and of each of
8	the plurality of previously categorized documents according to the hierarchical arrangement;
9	and
10	compare the semantic content of the uncategorized document with the
11	semantic content of each of the plurality of previously categorized documents to determine a
12	category for the uncategorized document.
1	38. The system recited in claim 37 wherein the categorization scheme
2	comprises a hierarchical categorization scheme.

2

1

2

1

2

1

2

- 39. The system recited in claim 37 wherein the engine is configured to compare the semantic content by producing a matching score between the uncategorized document and each of the plurality of previously categorized documents.
- 40. The system recited in claim 39 wherein the engine is configured to produce the matching score by determining a distance within the hierarchical arrangement.
- 1 41. The system recited in claim 40 wherein determining the distance 2 comprises accounting for a qualia relationship between types in the hierarchical arrangement.
  - 42. The system recited in claim 40 wherein the matching score is filtered to increase the importance of a smaller distance relative to a larger distance.
  - 43. The system recited in claim 37 wherein the engine is in communication with the internet.